Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence

Ulintz, Peter J ; Bodenmiller, Bernd ; Andrews, Philip C ; Aebersold, Ruedi ; Nesvizhskii, Alexey I

Abstract: Improvements in ion trap instrumentation have made n-dimensional mass spectrometry more practical. The overall goal of the study was to describe a model for making use of MS(2) and MS(3) information in mass spectrometry experiments. We present a statistical model for adjusting peptide identification probabilities based on the combined information obtained by coupling peptide assignments of consecutive MS(2) and MS(3) spectra. Using two data sets, a mixture of known proteins and a complex phosphopeptide-enriched sample, we demonstrate an increase in discriminating power of the adjusted probabilities compared with models using MS(2) or MS(3) data only. This work also addresses the overall value of generating MS(3) data as compared with an MS(2)-only approach with a focus on the analysis of phosphopeptide data.

DOI: https://doi.org/10.1074/mcp.M700128-MCP200

MCP Papers in Press. Published on September 13, 2007 as Manuscript M700128-MCP200

Investigating MS^2-MS^3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence

Peter J. Ulintz[1,2], Bernd Bodenmiller[3], Philip C. Andrews[1], Ruedi Aebersold[3,4,5], and

Alexey I. Nesvizhskii[6,2*]

[1]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan, 48103, USA.

[2]Bioinformatics Program, University of Michigan, Ann Arbor, Michigan, 48103, USA.

[3]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland.

[4]Institute for Systems Biology, Seattle, Washington, USA.

[5]Faculty of Science, University of Zurich, Switzerland.

[6]Department of Pathology, University of Michigan, Ann Arbor, Michigan, 48103, USA.

*** Corresponding author:**

Alexey I. Nesvizhskii

Department of Pathology, University of Michigan

4237 Medical Science I

Ann Arbor, MI, 48109

Email: nesvi@med.umich.edu

Tel: +1 734 764 3516

Running Title: Analysis of consecutive MS^2/MS^3 Spectra

Abbreviations

MS^2: Tandem Mass Spectrometry

MS/MS: Tandem Mass Spectrometry

MS^3: 3-Stage Mass Spectrometry (MS/MS/MS)

ICR: Ion Cyclotron Resonance

ROC: Receiver-Operator Characteristic

CID: Collision-induced Dissociation

FDR: False Discovery Rate

Summary

Improvements in ion trap instrumentation have made n-dimensional mass spectrometry more practical. The overall goal of the study is to describe a model for making use of MS^2 and MS^3 information in mass spectrometry experiments. We present a statistical model for adjusting peptide identification probabilities based on the combined information obtained by coupling peptide assignments of consecutive MS^2 and MS^3 spectra. Using two data sets, a mixture of known proteins and a complex phosphopeptide-enriched sample, we demonstrate an increase in discriminating power of the adjusted probabilities, compared to models using MS^2 or MS^3 data only. This work also addresses the overall value of generating MS^3 data as compared to an MS^2-only approach, with a focus on the analysis of phosphopeptide data.

Introduction

Advances in mass spectrometer design continue to propel proteomics research. One of the most widely used mass analyzers for protein work has historically been the ion trap, and a large proportion of the data from current mass spectrometry-based proteomics experiments are generated on such instruments. This trend continues with current generation 'linear trap' instruments that are characterized by increased ion capacity and thus improved resolution and sensitivity (1,2). Standard proteomics approaches are based on the predictable fragmentation of peptides in the collision cell of the mass spectrometer and the subsequent interpretation of the resulting spectra to infer amino acid sequence, referred to as tandem mass spectrometry (MS/MS or MS^2) (3-7). In practice, however, acquired MS/MS spectra are often noisy, contain only a small number of fragment ions due to incomplete peptide fragmentation, or reflect unanticipated instrumental or chemical artifacts. As a result, in a typical analysis of MS/MS spectra generated in a large scale experiment, only a small fraction

of the spectra can be successfully interpreted and assigned a peptide sequence with high confidence (8,9).

Newer instrumentation supports alternative techniques for data generation that have the potential to improve peptide and protein identification. One such technique is 3-stage mass spectrometry (MS^3), in which peptide ions in an ion trap or ICR mass spectrometer are subjected to an additional stage of isolation and fragmentation. The faster acquisition times of newer linear trap instruments such as the LTQ provide the option of collecting MS^3 spectra of abundant MS^2 peaks with overall cycle times similar to those of normal MS/MS^2 cycles on older 3D trap instruments. As a result, a number of researchers are choosing to routinely collect MS^3 spectra during LC-MS/MS runs which have the potential to provide additional information useful for peptide identification and characterization. This is deemed particularly important in the case of proteins identified by single peptides (10, 11) and for the analysis of phosphopeptides, the spectra of which are frequently dominated by a major fragment ion representing neutral loss of the phosphate group from the precursor peptide. Therefore, phosphopeptides have been analyzed by automated data-dependent triggering of MS^3 acquisition whenever the dominant neutral loss ion of the appropriate mass is detected in an MS^2 spectrum (12-14). Fragmentation of the neutral loss ion typically provides significantly increased structural information via increased peptide bond cleavage. Similar approaches may be applied to other major neutral loss ions (e.g. loss of 64 Da from peptides containing methionine sulfoxide) and to excessive prolyl- or aspartyl-directed fragmentation. MS^3 spectra have proven to be useful in top-down analysis as well, both for protein identification and for characterization of specific sites of post-translational modification. (15, 16)

Generally speaking, there are several ways of combining MS^2 and MS^3 spectra from the same peptide to improve peptide identification. One strategy involves integrating matching MS^2 and MS^3 spectra directly at the spectrum level, generating an "intersection spectrum" that contains only one type of ion, thus allowing simplified *de novo* sequencing of

the peptide. This approach has been described by Zhang and McElvain, who demonstrated the technique's usefulness in protein sequencing (17). Olsen and Mann describe a custom scoring algorithm for MS^3 spectra: their final score for a peptide is the product of the Mascot-generated MS^2 and the custom MS^3 score (11). In glycoproteomics, it is frequently the case that MS^2 and MS^3 provide complementary structural information on a glycopeptide: information on the structure of side-chain carbohydrate moieties is generally obtained from the MS^2 spectrum, while amino acid sequence information is more readily obtained in the MS^3 (18). In the top-down technique described by Zabrouskov et al (16), sequence tags are extracted from MS^3 spectra using a *de novo* algorithm and used to complement correlated MS^2 spectral data in a "hybrid" database search strategy, implemented in the ProSight PTM search engine (19).

Related to the problem of MS^2-MS^3 spectrum integration, *de novo* sequencing-based algorithms have been described for combining pairs of spectra corresponding to unmodified and modified versions of the same peptide, or pairs of spectra corresponding to the same peptide tagged with a light or heavy version of a labeling reagent (20-23). However, while *de novo* sequencing approaches are promising, no computational tools are currently available that can be robustly applied in a high throughput environment. As a result, analysis of MS^2 and MS^3 data is still largely carried out with a conventional database search approach using commercially available programs such as SEQUEST, MASCOT, SpectrumMill, Phenyx, Paragon, or open source programs X! Tandem, OMMSA, Inspect, or ProbID (24-29).

While all existing database search tools can be used to identify peptides from both MS^2 and MS^3 spectra, automated analysis of those different types of spectra may not be identical. This often leads to the requirement that MS^2 and MS^3 spectra be separated for processing. The main reason for this is that the measured precursor mass associated with MS^3 spectra will not always correspond to the mass of an appropriate database peptide calculated using the same conventional rules that are applied in the case of MS^2 spectra. For

example, in phosphopeptide analyses variable modifications of -18 Da due to loss of phosphoric acid from S or T residues need to be specified for MS^3, while the normal +80 Da phosphorylation modification on S, T, and Y are used for MS^2. It is computationally inefficient, and an unnecessary source of false positive identifications, to perform a combined search which permits both the -18 Da loss for MS^2 spectra and the +80 Da addition for MS^3 spectra.

Searching MS^3 spectra separately from their parent MS^2 spectra essentially decouples the two sets of scans. Intuitively, if analysis of successive MS^2 and MS^3 scans results in matching peptide sequences, there is an increased confidence in both identifications. The work described here attempts to provide a general, statistically sound assessment of the confidence achieved by combining the search results of MS^2 and MS^3 spectra from the same peptide. In contrast to aforementioned work, we assume a workflow in which the MS^2 and MS^3 spectra are searched independently using a common search engine (namely, SEQUEST in this work) and are independently statistically validated using PeptideProphet. We then re-couple matching consecutive MS^2 and MS^3 scans and adjust the peptide probabilities initially computed by PeptideProphet to account for the new "linked" MS^2-MS^3 information. We describe a model that produces an adjusted probability of peptide identification and demonstrate, using a data set of MS^2 and MS^3 spectra generated using a control protein mixture, that such a correction can be used to better discriminate between correct and incorrect database search results. We also investigate ways to combine the adjusted MS^2 and MS^3 probabilities to compute a single confidence measure for their corresponding unique peptide. We then further demonstrate the utility of our method using a phosphopeptide-enriched data set generated from *D. melanogaster* samples on an LTQ linear ion trap instrument. Finally, we compare runs in which both MS^2 and MS^3 spectra are generated with an MS^2-only method to address the overall benefit of generating MS^3 data.

Experimental Procedures

Sample Preparation and Mass Spectrometry

Two experimental data sets of MS/MS spectra were used in this work to evaluate the statistical model and to investigate its utility in the analysis of phosphopeptide-enriched samples. All spectra were acquired using an electrospray ionization (ESI) linear ion trap tandem mass spectrometer (Thermo Electron's LTQ).

(1) Nine-Protein Mix ("9-Mix") sample. A mixture of nine commercially available protein standards-- P68082, myoglobin of Equus caballus (horse); P00698 Lysozyme C precursor of Gallus gallus (Chicken); Q29443 Serotransferrin precursor (Transferrin) of Bos Taurus; P18915 Carbonic anhydrase 6 precursor of Bos taurus (Bovine); P12763, Alpha-2-HS-glycoprotein precursor (Fetuin-A) of Bos taurus (Bovine); P02754 Beta-lactoglobulin precursor (Beta-LG) of Bos taurus (Bovine); P62894 Cytochrome C of Bos taurus (Bovine); P02666 Beta-casein precursor of Bos taurus (Bovine); P02769 Serum albumin precursor (BSA) of Bos taurus (Bovine)-- was digested using trypsin and the resulting peptide mixtures were purified using reverse phase chromatography prior to mass spectrometric analysis. For the analysis of the peptides using mass spectrometry see "Mass spectrometry". The final data set consisted of three LC-MS/MS runs, with 58081 MS/MS spectra in total.

(2) "Phosphopeptide sample". This sample is a trypsin-digested, IMAC-enriched *D. melanogaster* whole cell lysate. The preparation of the phosphopeptide samples is described in detail in Bodenmiller et al. (30). Several mass spectrometry analyses of this sample were conducted, both for analysis of performance of the probability model and to test the value of generating MS^3 data.

Mass spectrometry: An LTQ quadrupole linear ion trap mass spectrometer (ThermoElectron, San Jose, CA) was used with a HP 1100 solvent delivery system (Agilent, Polo Alto, CA) for the analysis of the *D. melanogaster* Kc167 cells cytosolic phosphoproteome. Peptides were

loaded on a capillary (BGB Analytik, Böckten, Switzerland) reverse-phase C_{18} column (75 μm i.d. and 11 cm of bed length with Magic C18 AQ 5 μm 200Å resin (Michrom BioResources, Auburn, CA, USA)), and then eluted from the capillary column at a flow rate of 200-300 nl/min to the mass spectrometer through an integrated electrospray emitter tip. Peptides were eluted for each analysis from 12% to 33% acetonitrile in which the ions were detected, isolated, and fragmented in a completely automated fashion. The exact settings for MS^n acquisition were as follows:

9 protein mix: In the first scan event, all peptides eluting from the column were recorded in MS mode. The most intense ion was selected for product ion spectrum (MS^2) in the second event. An MS^3 spectrum of the most intense peak in the MS^2 spectrum was automatically selected in the third scan event. The second and third events are then repeated two more times in the cycle, for the second and third most abundant MS^1 ions, for a total cycle of seven events. A threshold of 5,000 ion counts was used for triggering an MS^2 attempt. Wideband activation was enabled for all MS^2 and MS^3 scan events; MS^2 isolation width was set to 2.0 m/z and MS^3 isolation width was set to 4 m/z. For triggering an MS^3 event the most intense ion had to be above 50 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

Phosphopeptide sample: All peptides eluting from the column were recorded in MS mode in the first scan event. The most intense ion was selected for product ion spectrum (MS^2) in the second event. An MS^3 spectrum of the most intense peak in the MS^2 spectrum, which for the phosphopeptide containing sample is in most cases the neutral loss peak (of 98 Da) from a serine/threonine phosphopeptide, was automatically selected in the third scan event. These three events form one complete cycle. A threshold of 20,000 ion counts was used for triggering an MS^2 attempt. Wideband activation was enabled for all MS^2 and MS^3 scan events. MS^2 isolation width was set to 2 m/z and MS^3 isolation width was set to 3 m/z.

For triggering an MS^3 event the most intense ion had to be above 500 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

Phosphopeptide sample – additional data sets for comparison of MS^2-only with MS^2/MS^3 methods: For the MS^2/MS^3 data set the data-dependent MS^n spectra were acquired as follows: in the first scan event, all peptides eluting from the column were recorded in MS mode, and then the most intense ion was selected for product ion spectrum (MS^2) in the second event. In the third event a MS^3 spectrum was triggered specifically in the event of a phosphate neutral loss (-98 Da for singly, -49 Da for doubly and -32.66 Da for triply charged peptides) in the MS^2 event. The second and third events are then repeated two more times in the cycle, for the second and third most abundant MS^1 ions, for a total cycle of seven events. For the MS^2-only data set the data-dependent MS^n spectra were acquired as follows: in the first scan event, all peptides eluting from the column were recorded in MS mode, and then the three most intense ions were consecutively selected for product ion spectrum (MS^2) for a total cycle of four events. Further settings for these samples were: wideband activation was enabled for all MS^2 and MS^3 scan events, MS^2 isolation width was set to 2 m/z and MS^3 isolation width was set to 4 m/z. For triggering an MS^3 event in the MS^2/MS^3 data set the most intense ion had to be above 50 ion counts. No further restrictions were made for the selection of the MS^3 precursor.

Database Searching and Results Analysis

MzXML files were generated from ThermoFinnigan *.raw files using the ReAdW tool available in the TPP platform (31-33). MS^2 and MS^3 peaklist files in *.dta format were extracted separately from the mzXML files using mzXML2Other tool with the -level option[1]. For the 9-Mix data set, a custom fasta sequence file was constructed consisting of sequences

[1] http://tools.proteomecenter.org/software.php

corresponding to the proteins in the mixture and common contaminants appended to a reversed version of the IPI Human data set. Resulting *.dta files for the 9-Mix data set were searched with SEQUEST using the following parameters: peptide tolerance of 3.0 Da; b- and y-ion series; partial trypsin digestion, allowing for one missed cleavage site; a fixed modification of 57.02 was specified for Cysteine and a variable post-translational modification (PTM) of 16.0 to Methionine. MS^3 data sets were searched using identical parameters. Note, partial trypsin specificity is required for searching MS^3 spectra corresponding to the fragmentation of a selected y- or b-ion from the MS^2 spectrum. If sufficient computational resources are available, searching MS^2 spectra allowing for partially tryptic peptides can often be beneficial and result in additional identifications. However, doing so requires that the results are properly analyzed with a tool that accommodates tryptic termini information in the statistical model, such as PeptideProphet. In addition, a subset of MS^3 spectra from this data set was also searched allowing for the C-terminus variable modification of -18.0 Da to accommodate the possibility that the MS^3 precursor is a b-ion (11). The results indicated that including this modification does not significantly alter the overall performance; in fact, accommodating the variable modification decreases the number of identifications slightly (due to loss of a number of true peptide assignments because of increases in search space). Based on this, the C-terminal modification was not used in the final analysis of data presented in this manuscript. The resulting data set contained 76873 peptide assignments, counting 2+/3+ duplicates: 48921 MS^2 (554 singly charged, 24233 doubly charged, and 24134 triply charged), and 27952 MS^3 (4582, 11700, and 11670 singly, doubly, and triply charged, respectively). Note that because of the charge state ambiguity (in the case of low mass accuracy data such as the data sets used in this work, the charge state of a multiple charged peptide ion cannot be reliable determined), most of the multiply charged spectra were searched twice, assuming 2+ or 3+ charge state. Furthermore, due to a relatively

small number of singly charged MS^2 spectra, all such spectra were left out of the subsequent analysis.

The database for the phosphopeptide-enriched samples consisted of all *Drosophila melanogaster* sequences exported from the UniProt database (34), 26311 entries total, to which the reversed set of sequences was appended. Parameters for the MS^2 search were: peptide tolerance of 3.0 Da; partial trypsin digestion, one possible missed cleavage; fixed modification of 57.02 for Cysteine; variable modifications of 80 Da were specified for S, T, and Y; a maximum 4 PTMs per peptide. The MS^3 spectra were searched with the same set of parameters except that variable modifications of -18 Da on S and T (instead of +80 Da) were specified to accommodate loss of phosphoric acid leading to a dehydroalanine or dehydrobuyric acid, respectively. SEQUEST database searching for the primary phosphopeptide data set (excluding the MS^2/MS^3 to MS^2-only comparisons) resulted in 28865 peptide assignments, counting 2+/3+ duplicates: 16647 MS^2 (143 singly charged, 8483 doubly charged, and 8021 triply charged), and 12218 MS^3 (547, 5895, and 5776 singly, doubly, and triply charged, respectively).

The additional phosphopeptide-enriched data sets used for comparison of MS^2/MS^3 and MS^2-only methodologies consisted of the following number of peptide assignments following SEQUEST database searching- Run 1 (A07_5205): 4915 MS^2 assignments (95 singly charged, and 2410 each of doubly and charged), and 1897 MS^3 assignments (31 singly and 933 each doubly and triply charged); Run2 (A07_5206): 6450 MS^2 assignments (126 singly, 3162 doubly and triply charged); Run 3 (A07_5207): 4883 MS^2 assignments (103 singly charged, and 2390 each of doubly and charged), and 1879 MS^3 assignments (43 singly and 918 doubly and triply charged); and Run 4 (A07_5208): 6403 MS^2 assignments (159 singly, 3122 doubly and triply charged).

Processing of MS^2 and MS^3 search results

Search results for each LC-MS/MS run were generated by first producing an html results file using the out2summary tool, exporting one result file for each MS level, for each run: a total of six files for the 9-Mix data set and two files for the phospho data set. Html results were then converted into pepXML format (31) using Sequest2XML. PeptideProphet (32) was run on each result set, generating probability scores for each search result that are added to the pepXML documents.. For the phospho data sets, PeptideProphet was run with the "-l" option, which results in alternate processing of DeltaCn scores marked with '*': results for which the top and second-highest ranked peptide assignment to a spectrum have homologous sequences (>70% sequence identity). With this option on, PeptideProphet will use the Xcorr score of the first non-homologous lower scoring peptide match when computing DeltaCn score of the best scoring peptide. This option is beneficial in the event that the search returns several identical results that differ only by modification site for a sequence, as often occurs in phoshorylated peptide identifications.[2] Resulting files were parsed and processed to generate all matching statistics using a custom set of scripts implemented in Python. Certain subsets of data were also exported into a local Mysql database instance to facilitate generation of specific statistics.

Linking MS^2 and MS^3 scans and search results

The spectra in these experiments were generated in an interlaced manner, i.e. the scan cycle on the instrument followed the format: MS^1->MS^2->MS^3->MS^2->MS^3->MS^2->MS^3, or MS^1->MS^2->MS^3, with the MS^2 scans triggered in a data dependent manner from the MS^1, and the MS^3 scans triggered from the preceding MS^2. As a result, a set of "linked" MS^2/MS^3

[2] The default option in PeptideProphet is to set SEQUEST DeltaCn score to zero to reduce the probability that the best scoring peptide assignment to a spectrum is correct when the second best scoring peptide has high sequence homology.

scans were generated based on consecutive scan numbers. In the resulting data set, MS^2 scans with no consecutive MS^3 were retained and designated as linked, but as a link to a null MS^3 identification. MS^3 scans without preceding MS^2 scans should not occur physically, but do in these data for several reasons: namely, the corresponding MS^2 peaklists that produced no database search result are typically not reported. Also, some spectra containing only a few peaks may be filtered out by the data conversion software. The small number of instances in which these "orphaned" MS^3 scans are generated invariably result in incorrect peptide identifications and are eliminated from subsequent analysis.

Due to uncertainty with the charge state each multiply charged scan was searched twice (in both 2+ and 3+ charge state), resulting in multiple search results for each scan. Consideration needs to be given to potential links between MS^2 and MS^3 search results for any pair of scan numbers. A +1 MS^2 search result may only be linked to an MS^3 search result that is +1, and a +2 MS^2 scan may produce a link to a search result with either a +1 or +2 charge state. The double- and triple-charged SEQUEST search duplication, however, creates a situation in which a +3 MS^2 search result may produce two possible links to +2 and +3 MS^3 search results for any pair of scan numbers. After generating all possible links, one pair of search results amongst all possible pairs for any two scan numbers (designated as the "unique pair") is selected based on whether the sequences of the two peptide identifications composing a pair are matching. Matching is defined here as whether or not the sequences are equal, or whether one contains a subsequence of the other. For non-matching pairs, and scan sets with more than one pair with matching sequences, the match pair with the highest summed PeptideProphet probability is designated as the unique pair. A schematic of all matching possibilities and selection of a unique pair is shown in Supplementary Figure 1.

Results and Discussion

Overview of the probability adjustment method

The overall methodology for our approach is outlined in Figure 1. Data generated by the mass spectrometer are processed via the Trans-Proteomic Pipeline (TPP) following normal procedures and using SEQUEST, Mascot, or X! Tandem database search tools for peptide identification (the tools currently supported by TPP), up through generation of peptide probabilities from PeptideProphet (32). Analyses in this early stage of processing are conducted separately for MS^2 and MS^3 data. To calculate an adjusted probability for all assignments, successive scans must be linked as described in the Methods section. The multiple potential matches resulting from the charge state ambiguity are reduced in the processing, retaining only the most probable matching pair for any two scan numbers.

Based on the sequence of the highest scoring peptide produced by the database search tool for each scan, consecutive MS^2/MS^3 pairs may then be classified as to whether or not they match the same peptide sequence. This classification forms the basis for the adjusted probability score (see below), which functions to reward assignments with matching sequences. Only the top-ranked peptide sequence for each spectrum is used in this analysis; accommodation of lower ranking results, while potentially useful, is not considered for simplicity. The result of the probability correction procedure is a data set of linked MS^2 and MS^3 peptide identifications with adjusted probability scores.

Linking MS^2 and MS^3 data: a case study of the 9-Mix data set

This analysis is carried out using a mixture of purified proteins (9-protein mix data set), in which it is possible to confidently label peptide identifications as ‘correct’ or ‘incorrect’. Because this data set was searched against a database consisting of the sequences of the mixture proteins appended with a much larger reversed human protein sequence

database, each spectrum could be assigned a correctness label based on whether the top SEQUEST hit for the spectrum was to one of the known protein entries. The method used was simply to label as incorrect any assignment of a peptide from a known incorrect database entry (reversed human protein sequence entries in this case), whereas all assignments of peptides to one of the sample proteins can be considered correct (32).

The procedure begins by linking consecutive MS^2 and MS^3 scans using their scan numbers. The summary statistics of applying the linking procedure to the 9-Mix data set are shown in Table 1. Due to the uncertainty in the precursor charge state for LTQ spectra, for any pair of consecutive MS^2/MS^3 scan numbers, there may be one or two SEQUEST search results generated for each MS level, as described in Experimental Procedures. Consequently, an MS^2 search result may be linked to more than one MS^3 search result. For the 9-Mix data set, there are 16140 unique linked pairs in which the MS^3 is not null. Amongst these, eighty nine have MS^2/MS^3 charge states of +1/+1, eight of which match "correct" protein sequences in the database (either one or both of the sequences match). For doubly-charged MS^2 pairs, 3761 are +2/+2 and 4043 are +2/+1, of which 878 and 2020 are correct, respectively. For triply-charged MS^2, +3/+3: 4020 pairs, 631 correct; +3/+2: 3777 pairs, 1177 correct; and +3/+1: 450 pairs, 111 of which are correct. In all, linked pairs in which the MS^3 has one less charge than the MS^2 are more likely to be correct. However, linked pairs for which the MS^3 is the same charge state as MS^2 account for 36% of the correct identifications.

Neutral loss of amino acids from the N- and C-termini is a common phenomenon and has been described previously (35, 36). Selecting linked pairs in which both MS^2 and MS^3 sequences are labeled correct and of the same charge state (+1/+1, +2/+2, and +3/+3) allows us to identify examples of amino acid neutral loss. Our data confirms the conventional rules for amino acid neutral loss described in the literature. Virtually all examples correspond to N-terminal loss of 1-4 amino acid residues, most frequently N-terminal to a proline. 276 out of 323 of the occurrences are doubly-charged, three are singly-charged, and the remaining

forty-four triply-charged. Most examples occur multiple times: in all there are one, thirty-four, and nine unique neutral loss sequence examples for the singly-, doubly-, and triply-charged cases, respectively. These examples are provided in Supplementary Table 1.

After linking consecutive scans and selecting a unique linked pair, the peptide assignments are binned into sequence match categories dependent on whether a consecutive scan exists, and if so, whether the top-scoring SEQUEST sequence result of the successive scans match (Table 1b). Sequence match categories (referred to as Match categories, or simply 'Match' later in the text) are defined as follows: 0) no consecutive scan; 1) consecutive scans, but MS^2 and MS^3 sequences do not match; 2) consecutive scans, MS^3 sequence is a subset of the MS^2 sequence; 3) consecutive scans, MS^3 sequence identical to the MS^2 sequence; and 4) consecutive scans, MS^2 sequence a subset of MS^3 sequence. In the data set of unique pairs, 69% of all MS^2 spectra produced consecutive MS^3 spectra (16140). Out of those consecutive pairs, 1458 (9%) had matching sequences in which the MS^3 sequence was a subset of the MS^2 sequence. 116 MS^3 spectra were orphaned because they did not have a preceding MS^2 scan, and were discounted. We note that there were no instances of identical sequence matches between MS^2 and MS^3 top-scoring hits in the 9-Mix data set, as may occur for neutral-ion events in which only a side-chain moiety is lost from the otherwise intact peptide backbone (e.g. a phosphate). These losses are observed in other similar data sets, however, and do occur in the phospho-enriched data sets described later.

For a small number of linked pairs, the top-scoring MS^3 sequence appears to be a superset of the MS^2 sequence, binned as sequence match category 4. Clearly such pairs are not physically possible. Detailed analysis indicated that that most of those cases can be explained as resulting from misidentification of the true peptide sequence from either MS^2 or MS^3 scan. For example, in some of these instances, the sequence corresponding to the +2 MS^2 is a subsequence of both the +3 MS^2 sequence and the +2 MS^3 sequence, with the +2/+2 MS^2/MS^3 pair selected as the unique pair. In those cases, the peptide assignment to the +3

MS^3 peaklist (with +3 being the true charge state of the peptide ion) scored lower than the assignment of a shorter peptide (a subsequence of the true peptide) to the +2 MS^3 peaklist. Other examples involved cases of a high scoring assignment of a longer partially tryptic peptide sequence when the true peptide was a post-translationally modified tryptic peptide missed due to the restricted nature of the database search. Similarly, several cases were observed where an MS^3 scan acquired on a doubly charged b-ion fragment from the parent MS^2 spectrum resulted in a match of a longer sequence to the +3 MS^3 peaklist, and no match in the case of the correct +2 charge state. In any event, as can be seen from Table 1b, match category 4 represents a small number of special case instances. For simplicity of articulation, this category is dropped from subsequent analysis.

Using the labeling of the data, the accuracies and sensitivities of the probability calculations could be determined. Towards this end, each linked pair of spectra can also be assigned a truth category based on the correctness of the peptide assignments to the MS^2 and MS^3 scans. The truth category is a label indicating whether neither, both, or which one of the matching scans has a 'correct' label. The total numbers of scans in each truth category are shown in Table 1c. The number of unique pairs of search results in which both sequences were correctly assigned is 1509, corresponding to 6.4% of the total number of unique pairs of scans. A greater number of linked pairs (3316 total, 14.2%) have either the MS^2 only assigned correctly (2029), or only the MS^3 (1287).

When comparing the counts in the sequence match category bins (Table 1b) with the truth category bins (Table 1c), there appear to be several (thirty four) more +/+ truth matches than expected from the number of entries in the sequence match bin categories 2 and 4. These entries are the result of sequence match category 1 entries contributing to the +/+ truth bin. There are a number of cases in which the top-scoring MS^2 and MS^3 sequences both match one of the sample mix proteins, but the proteins are different or the match is to different peptides from the same protein. Most of the instances are examples of the latter

case: a homologous sequence in the protein TRFE_BOVIN results in two different peptides (CLMEGAGDVAFVK and KGDVAFVK) being identified in the joined pairs. One of the commercially obtained proteins in the mixture, TRFE_BOVIN, was also contaminated with the homologous TRFL_BOVIN, which exhibits 59% sequence identity. As a result, homologous but not identical peptide sequences between the two proteins are identified in the joined pairs. For four cases, however, even though both MS^2 and MS^3 identifications in the pair are labeled correct in that individually their sequences match one of the sample proteins, there is no similarity between the matching sequences. These can be considered as chance matches to one of the sample mix proteins incorrectly labeled as correct (the observed number of such chance matches is consistent with the expected number given the relative sizes of the 9-Mix and the reversed Human protein sequence database). In all of such cases, either the MS^2 or the MS^3 was a high-probability result with the other joined probability very low.

Probability adjustment calculation

In automated analysis of mass spectrometry data, one of the most important tasks is the calculation of accurate and discriminative confidence measures for each peptide assignment to a spectrum produced by a database search tool. Towards that end, we seek to calculate a correction to the probability score that accommodates the increase in confidence resulting from matching MS^2 and MS^3 spectra. The fact that matched consecutive MS^2 and MS^3 spectra are more likely to be correct forms the basis for adjusting the probabilities of these spectra.

Calculation of probabilities for each peptide assignment in the data set, performed independently for MS^2 and MS^3 data, represents the starting point in this analysis. PeptideProphet computes a probability for a peptide, designated here as $p(+|D)$, by using the mixture model EM algorithm to model the distributions of various discriminant spectrum-

level parameters, collectively represented here as D. The spectrum-level information D typically includes the discriminant database search score (a linear combination of the renormalized search scores reported by the database search tool used), the number of termini consistent with the specificity of the enzyme used to digest proteins, the number of missed internal cleavage sites, and the difference between the measured and the calculated precursor ion mass. In certain cases, additional parameters are included in the model such as the peptide pI value (37), or the presence of certain residues or sequence motifs in the sequence of the assigned peptide (e.g., the presence of a cysteine in the case of ICAT experiments, or NxS/T motif in the case of experiments employing glycopeptide-enrichment strategies). PeptideProphet probabilities are reasonably accurate for both MS^2 and MS^3 spectra. A plot displaying probability accuracies of PeptideProphet results for the 9-Mix data is provided in Supplementary Figure 2.

The approach used to accommodate the additional sequence matching information is similar to the method described in (33) for adjusting probabilities to account for additional protein level information using the number of sibling peptides (NSP). The MS^2/MS^3 sequence match information is not available at the initial data analysis step, but can be used to adjust the initial probabilities *p(+|D)* after linking the corresponding MS^2 and MS^3 scans. Again, the adjustment is performed separately for MS^2 and MS^3 level data. Given the sequence match category (Match) assignments for all linked spectra, the adjusted probability of a linked peptide assignment from a certain sequence match category, *p(+|D, Match)*, may be calculated as:

$$p(+\,|\,D,Match) = \frac{p(+\,|\,D)\,p(Match\,|\,+)}{p(+\,|\,D)\,p(Match\,|\,+) + p(-\,|\,D)\,p(Match\,|\,-)} \qquad (1)$$

where *p*(*Match*|+) and *p*(*Match*|-) represent the empirically derived probabilities of observing a peptide assignment in each Match category among all (MS^2 or MS^3) correct and incorrect peptide assignments in the data set, respectively. Note that this calculation assumes that the

information derived from linking consecutive scans is independent of the identification information generated by a search engine. This is largely true. Normalized PeptideProphet SEQUEST discriminant score distributions for correct and incorrect peptide assignments to MS^2 spectra of doubly charged precursor ions, plotted separately for peptide assignments to MS^2 spectra belonging to different Match categories, are shown in Supplementary Figure 3; score distributions are similar for all values of Match parameter, justifying the assumption of the independence between the discriminant database search score and Match parameter.

The probability distribution $p(Match|+)$ may be calculated for each match category k as follows:

$$p(Match \mid +) = \frac{1}{Np(+)} \sum_{\{i|Match_i \in k\}} p(+ \mid D_i, Match_i) \quad (2)$$

where N is the total number of (MS^2 or MS^3) peptide assignments in the data set, and the sum is over all peptides i in each Match category. The term $p(Match|-)$ is calculated in a similar way. The overall proportion $p(+)$ of correct assignment in the data set may be calculated as:

$$p(+) = \frac{1}{N} \sum_i p(+ \mid D_i, Match_i) \quad (3)$$

The probabilities in Eq. 1, and the Match parameter distributions in Eq. 2, can be determined by starting with the initial PeptideProphet probability for each assignments, $p(+|D_i)$ and the overall proportion, $p(+)$. The probabilities and Match distributions can then be updated in an iterative manner. However, a single iteration was deemed to be sufficient for the data set used in this work.

Application of the probability adjustment method to the 9-Mix data set

Table 2 lists $p(Match|+)$ and $p(Match|-)$ distributions calculated using Eq. 2 for the 9-Mix data set for both MS^2 and MS^3 scans. It can be seen that, in the case of MS^2 spectra, a larger fraction of incorrect assignments have no consecutive matching scan. For all instances,

the most likely sequence match category is category 1, corresponding to the case in which consecutive scans occur but with no matching sequence. This is perhaps intuitive in the sense that it might frequently be the case that either the MS^2 or the MS^3 will produce an identifiable sequence, but not both. The most obvious discriminating measure is the fact that for 30% of the correctly assigned MS^2 spectra (the top row in the table), the linked MS^3 spectrum was assigned a peptide sequence that is a subset of the MS^2 sequence, as opposed to a 5% incidence for incorrect MS^2 identifications. If sequence matches are observed, identifications are thus much more likely to be correct; the same argument applies for MS^3 scans preceded by MS^2 scans. Also noteworthy is the fact that for match category 1 pairs, the probability of a correct identification is less than the probability of an incorrect identification. This will result in a probability penalty for consecutively linked scans without matching sequences. The penalty is small in this case, much smaller than the boost due to a consecutive matching scan, but is nevertheless an effect of the model.

It should be noted that in addition to classifying peptide match pairs into bins as a function of sequence matching, they can also be classified into various precursor charge state pairs. Significant differences exist between the precursor charge state distributions of correct and incorrect matches. An expansion of the sequence match category probabilities into charge category bins is provided in Supplemental Figure 4 for each of the four posterior Match probability distributions of Table 2, as well as total counts of the number of matches falling into each bin for the 9-Mix data set. The charge state information would likely provide additional discriminative power. However, further subclassification of the data into charge state pairs requires larger amount of data and complicates the model. Thus, the charge state information has not been utilized in the model at this time.

An example of the probability adjustment procedure described above is illustrated in Figure 2a using a pair of matching scans from the 9-Mix data set. MS^2 spectrum A06_7233_c.18651.18651 is first paired to MS^3 spectrum A06_7233_c_18652.18652 by

consecutive scan number. MS^2 assigned peptide sequence TLNFNAEGEPELLMLANWRPAQPLK is then compared to MS^3 sequence GEPELLMLANWRPAQPLK. Since the MS^3 sequence represents a fragment of the MS^2 sequence, the linked pair is assigned to sequence match category 2. The adjusted probabilities are then calculated for each spectrum using Eq 1. In this instance, the initial PeptideProphet probability of 0.712 is adjusted to 0.995 for the MS^2 spectrum, and 0.832 to 0.989 for the MS^3. A combined probability may then optionally be calculated for the linked pair as a new discriminating measure, as discussed later in the text.

Also indicated in Figure 2 are examples of fragmentation patterns from other charge state pairs. These examples are provided here to illustrate both differences in the relative extent of fragmentation that can occur as a function of charge and also the presence of redundant ions appearing in both the MS^2 and MS^3 spectra. Panels 2b – 2d contain examples from the phospho data set, specific features of which will be discussed in more detail later in the paper. It should be noted that many identical ions can be observed between matching MS^2 and MS^3 spectra.

In the development of the model, several (match category 2) cases were observed where both paired spectra had a low initial probability of being correct, but their probabilities became intermediate or even high values after adjustment. For example, the initial probabilities for peptide assignments to linked scans A06_7232_c.4362.4362.3 (MS^2 scan), and A06_7231_c.4363.4363.2 (MS^3 scan) of 0.077 and 0.319 would get boosted to 0.827 and 0.830, respectively, if the probabilities were adjusted using the Match parameter distributions shown in Table 2. Boosting such low probability assignments may be undesirable regardless of their match category. To address this, several approaches were investigated, including introduction of probability-dependent match categories. A very simple constraint that worked well in the case of the 9-Mix data set was to avoid any probability adjustment for Category 2

matches if both initial MS^2 and MS^3 probabilities were below a specified threshold, 0.5 in the case of these data. This was an optional feature that was investigated using the 9-Mix data set but not utilized for the phosphopeptide data sets, as it was deemed a minor adjustment that did not significantly affect the overall results; specifically, the number of entries in the 9-mix data set that were affected by this exception was only 24 out of a total 23367 unique matches

The improved discriminatory power of the adjusted probabilities, calculated using the $p(Match|+)$ and $p(Match|-)$ distributions shown in Table 2 (after the empirical correction described above), is indicated in Figure 3, which shows Receiver-Operator Curves (ROC) for the data. The performance of the model is evaluated separately for MS^2 and MS^3 spectra. The false positive error rate is plotted as a function of the sensitivity attainable by selecting a variable probability threshold. Sensitivity in this case is defined as the ratio of the number of correct peptide assignments to MS^2 (Figure 3a) or MS^3 scans (Figure 3b) with a probability greater than or equal to a specific probability threshold and the total number of correct assignments to MS^2 (4870) or MS^3 (1256) spectra, respectively. Similarly, the false positive error rate is calculated as the fraction of incorrect matches in the total number of spectra above each probability threshold. Note that there is redundancy between the MS^2 and MS^3 peptide assignments, so summing the total possible number of correct peptide identifications from both MS^2 and MS^3 scans would not reflect the total number of unique identifications.

For both the MS^2 and MS^3 scans, the adjusted probability provides a better performance profile, achieving greater sensitivity at an equivalent error rate as compared to the initial data. For example, at a 0.9 probability threshold, the initial MS^2 probability results in the selection of 4072 correct peptide assignments at the expense of 67 incorrect ones. Using the adjusted probabilities, selecting the same number of correct identifications results in only 38 incorrect peptide assignments.. The improvement in MS^3 discrimination is even more pronounced, especially in the optimal region of the curve. Using initial probabilities, 1350 correct and 19 incorrect assignments to MS^3 spectra pass the 0.9 threshold. Using the

adjusted probabilities, it becomes possible to select the same number of correct peptide assignments with the inclusion of only one false positive.

Combining MS^2 and MS^3 probabilities

The result of the probability adjustment procedure described above is now two adjusted probabilities for each unique linked pair of scans, one each for MS^2 and MS^3. Possibilities for best utilizing both of these scores in selection of correct and incorrect identifications are now explored. Ideally, a combined scoring approach would provide a greater discriminatory power for selecting correct and incorrect identifications than a subsequent counting of unique matches based on MS^2 and MS^3 taken individually. Two possibilities for utilizing both scores are examined:

$$P_{comb} = 1 - (1 - p_{MS^2})(1 - p_{MS^3}) \tag{4a}$$

$$P_{\max} = \max\left\{p_{MS^2}, p_{MS^3}\right\} \tag{4b}$$

where p_{MS2} and p_{MS3} are the adjusted probabilities for the MS^2 and MS^3 scans, respectively, for the same linked pair. The first option is appropriate when the two probabilities can be considered independent, and has been utilized (in a different context, i.e., for combining the evidence from different peptides) for the protein identification problem (33, 38). P_{comb} reflects the probability that at least one of the two peptide assignments, either to the MS^2 or to the MS^3 spectrum, is correct. However, it is obvious that MS^2 and MS^3 spectra, and therefore the probability scores p_{MS2} and p_{MS3} of those spectra, are not fully independent measurements of a peptide in that identical ions will be measured in both spectra. An alternative approach is to select the assignment with the highest probability, P_{max}, thus reducing the likelihood of possible overestimation of the final probability. P_{max} has been used in other similar situations, e.g. in selecting amongst several alternative equivalent peptides (assignments of

the same peptide to multiple MS/MS spectra) in the ProteinProphet protein probability score (33), and in Mascot protein-level scoring (24).

Figures 4a and 4b show the results of counting the number of correct peptide assignments above specified probability thresholds, utilizing all possible scores calculated for a linked pair as the discriminating measure: initial MS^2, initial MS^3, adjusted MS^2, adjusted MS^3, P_{max}, and P_{comb}. Displayed are the results on the set of all unique linked pairs. A comparison of the initial and adjusted probability results for MS^2 and MS^3 again demonstrates an increase in the number of selectable correct peptide assignments at any probability threshold as a result of the probability adjustment. Both P_{max} and P_{comb} scores perform similarly, and provide improved discrimination as compared to the individual measures. Obviously, the primary reason for the performance increase is the fact that the combined score permits the possibility of selecting either the MS^2 or the MS^3 for any linked pair, thus permitting a pair to be selected as correct if either probability is above threshold. At the 99% probability threshold, for example, the adjusted MS^2, adjusted MS^3, P_{max} and P_{comb} probabilities correspond to 3141, 1050, 3775, and 3807 correct peptide identifications, respectively. Figure 4c provides a measure of the rate of false-positives on these data for the most interesting thresholds. The same performance trends are evident: including roughly 40 false positives, specifically 40, 41, 39, and 39 for adjusted MS^2, adjusted MS^3, P_{max}, and P_{comb} measures, respectively, results in selection of 1806, 4139, 4594, and 4762 correct identifications. In all, P_{comb} provides the most discriminative measure.

In addition to analyzing the discriminative power of computed probabilities, one must also assess their accuracy. Probability accuracy plots for the adjusted and combined measures are shown in Fig 4d. The adjusted probability scores still provide an accurate representation of true probabilities and fit the 45° line well. The P_{comb} and P_{max} measures perform similarly well. Interestingly, P_{comb} does not overestimate probabilities as one might expect given the

dependence of MS^2 and MS^3 level spectra on this data set. Additional analysis would be necessary to determine if this is a general characteristic.

Phosphopeptide data set results

One of the main motivating factors in collecting MS^2/MS^3 data is to increase the confidence levels and the total number of phosphopeptide identifications. The identification of phosphopeptides from MS^2 spectra is challenging because spectra recorded using an ion trap mass spectrometer often exhibit one or more dominant neutral loss peaks of 98 Da, whereas the occurrence and intensity of the other fragment ions (containing peptide sequence information) may be impaired. To investigate potential improvement in discrimination as a result of the probability adjustment on a phosphopeptide-enriched data set, a data set of MS spectra from a single LTQ injection of an IMAC-enriched *D. melanogaster* sample was selected for detailed analysis in this work. The data were acquired in a data-dependent mode, with MS^3 scans triggered for the most abundant peak of the MS^2 spectra which in the case of this sample mostly corresponds to the neutral loss peaks: -98.00 (-116.00), -49.00 (-58.00), -32.60 (-36.66) Da from the precursor, as explained in the Experimental Procedures section. Since the sample in this case is a complex protein mixture, a precise labeling of peptide identifications as 'correct' or 'incorrect' is not possible. Instead, only the composite false discovery rates (FDR) (a single measure for each filtering threshold) can be estimated by counting the number of matches to reversed sequences.

The methodology for generating adjusted probability scores for this data set is analogous to the 9-Mix data set. Top-scoring MS^2 and MS^3 SEQUEST peptide assignments are linked based on consecutive scan numbers, and the top-scoring pair for consecutive scans is selected. Note that if MS^3 spectra are triggered based on neutral loss peaks, charge state ambiguity between matching pairs can potentially be reduced. This fact is not exploited in our analysis; rather, we maintain the same procedure for allowing all possible charge pairs in

a match. The match pairs are then classified into sequence match categories as described above. The same four sequence match categories are used: 0: no consecutive match; 1: consecutive match but no matching sequence; 2: matching sequences with MS^3 sequence a subset of MS^2 sequence; and 3: matching sequences with MS^3 sequence identical to MS^2. In this data set, there were only two instances of scans that would correspond to the sequence match category 4: matching sequences with MS^2 sequence a subset of MS^3 sequence. Again, this category was eliminated for simplicity. We note that the additional constraints imposed by the data-dependent triggering of these data and the resultant database searching provisions would allow us to generate additional useful sequence match categories, corresponding to whether the site of modification of a match is identical between the two sequences. We observed a number of instances in these data where the sequences matched but the sites of modification of the match did not, indicating ambiguity in the localization of the modified residues. A larger data set would allow a more rigorous analysis of these types of results (39, 40).

The number of search results generated for this data set is shown in Table 3a. SEQUEST searching produced 16647 and 12218 results for the MS^2 and MS^3 data sets, respectively, corresponding to 7547 unique matching pairs of searched results. Counts for the four sequence match categories are shown in Table 3b. Most significant is the fact that the sequence match category corresponding to neutral loss-only pairs (match category three) is no longer null; rather it is the more abundant category amongst the two representing matching sequences with 313 unique matches.

Corresponding posterior probabilities were calculated for the sequence match categories, and then used to calculate the final adjusted probability for each unique pair. These numbers are shown in Table 3c. The frequencies of observing a correct or incorrect assignment to an MS^2 scan with no matching MS^3 sequence (match category one) are relatively close; only a small probability correction occurs for these instances. MS^3 category

one probabilities are penalized, as are MS^2 instances that lack a corresponding MS^3 result. A probability boost is received for pairs in categories two and three, with a greater correction given to the latter.

Although a true sensitivity measure for these data is impossible, it is possible to evaluate the relative performance of the various probability measures by examining the number of reversed database matches. The decoy database method is increasingly being used as an effective means of estimating false positive rates in database searching when other methods of error rates estimation cannot be readily performed (41, 42). At any given probability threshold, the number of matches to reversed sequences can be calculated and compared to the total number of peptide assignments above that threshold to derive an estimate of the FDR (42). A measure of the performance of the various model probabilities on these data is shown in Figure 5a. The figure plots the estimated number of correct identifications as a function of FDR. These data are generated by ranking all peptide assignments in order of decreasing probability. The number of assignments of peptides from the forward database (n_f) having a probability equal or greater than the probability of the n^{th} top-ranking reverse entry (n_r) is counted, and the estimated false discovery rate is determined as n_r/n_f. The estimated number of correct assignments is similarly measured as $n_f - n_r$. This analysis is done separately for each of the initial and adjusted probability measures: MS^2 and MS^3 initial and adjusted, as well as the combined probability measures P_{comb} and P_{max}. A version of these data in table form is provided in Supplemental Table 2, which presents estimated false positive percentages and number of forward match counts for inclusion of one, two, five, ten, fifty, and one hundred reversed matches, as well as the number of those forward entries that are identified as containing phosphorylation sites.

As can be seen from Figure 5a, at equivalent false discovery rates, the adjusted probability measures for MS^2 and MS^3 data provide a small but distinguishable improvement in the number of correct entries that can be selected, particularly for MS^3. The bigger benefit

of course comes with the combined P_{comb} and P_{max} scores, which provide a much higher selection rate of forward matches than the initial MS^2 and MS^3 probabilities.. For example, by filtering the data using P_{max} instead of the initial MS^2 probability it becomes possible to extract 203 more forward matching identifications without allowing any reverse database matches (1703 peptide identifications vs. 1499). At a roughly 5% FDR, the initial MS^2 probability estimates 1893 correct peptides whereas the P_{max} measure selects 2093. It is interesting that P_{comb} is much more discriminative than the P_{max} probability measure on these data, selecting 2328 correct peptides at the 5% FDR. Overall, the acquisition of MS^3 spectra does appear to increase the total number of phosphopeptide identifications by 10-25% in this data set, depending on the specific combined probability score used for comparison.

The results discussed above for this sample have focused on the total number of identifications, the majority of which are phosphopeptides. An equivalent plot of the results, but including only ranked non-phosphorylated identifications from the phosphopeptide data set, is shown in Figure 5b. In general, the same trends can be seen; the model improves the assignment scores of unmodified peptides as well.

Example MS^2 and MS^3 spectra from the phosphopeptide data set

In order to understand the underlying reasons for improved identification confidence, it is informative to briefly revisit the example shown in Figure 2. These spectra are representative illustrations of matched MS^2 and MS^3 phospopeptide spectra of various precursor charge states. Several spectral features are of interest. Figure 2b shows an example of a +2/+1 match pair. The threonine in position three of the sequence matching the MS^2 spectrum is phosphorylated. The large y_{12} peak corresponding to a fragmentation n-terminal to a double proline was selected by the instrument for MS^3. This is a general characteristic of the singly charge spectra corresponding to correct identifications in these data: the majority are proline-directed, with a Pro identified in the first position. Although the fragmentation is

reasonable in this MS^3 spectrum, a large fraction of singly-charged spectra exhibit poor fragmentation with one or two major peaks corresponding to Pro, Asp or occasionally Glu cleavage dominating. This is not surprising due to the relatively low energy imparted to singly-charged ions via collision-induced dissociation (CID) in a trap instrument; typically the most facile fragments are the most readily observable. As can be seen, many of the same ions occur in both spectra. However, the shorter sequence and the absence of the phosphorylated residue in the MS^3 simplify the spectrum and increases confidence in the identification. Figure 2c shows a +3/+1 phosphopeptide example. +3/+1 instances are rarer than the +2/+1 (see Supplementary Figure 3), and the same trends occur. The MS^3 spectrum shown is a proline-directed fragmentation event, with Asp-directed fragmentation peaks dominating the spectrum.

Figure 2d is an example of a +2/+2 phosphopeptide ion. The peak selected for MS^3 corresponds to the doubly-charged y_{13} peak with a -98 Da loss of the phosphate moiety. Although many identical ions are identified in both spectra, there is a significant difference in the fragmentation pattern, with several ions observable in MS^3 which are not readily observable in the MS^2.

Data set dependence of probability adjustment

Since the two primary data sets used in this work differ significantly in terms of sample complexity, it is also informative to compare these two data sets with respect to the MS^2/MS^3 matching statistics and the degree to which the initial peptide probabilities are adjusted to account for the sequence match information. The Match parameter distributions *p(Match|+)* and *p(Match|-)* vary between the data sets, reflecting the differences in the sample complexity and data set size. This is illustrated in Figure 6, which plots the logarithm of the ratio *p(Match|+)/p(Match|–)* for each match category *k* for both data sets. A ratio greater than

1 (log ratio greater than 0) indicates the region where the probabilities are boosted after adjustment for Match information, whereas a ratio less than 1 (log ratio below 0) indicates that the Match adjustment reduces the probability that a peptide assignment is correct. While the overall trend is similar for both data sets, significant differences exist in the amount of adjustment. For example, the penalty applied to a peptide assignment to a MS^2 spectrum with no subsequent MS^3 spectrum (match category 0) is approximately twice as high in the case of the phosphopeptide enriched data set than in the 9-Mix data set. On the other hand, the amount of probability boost for peptide assignments in the Match=2 category is higher in the case of the 9-Mix data set. A better understanding of these results requires analysis of the MS^2-MS^3 linking statistics for a larger data set. However, it is clear that the amount of probability adjustment in each sequence match category is data set-dependent. Thus, it is advantageous to use statistical methods for combining MS^2 and MS^3-level data that can learn the appropriate amount of probability adjustment from the data itself, such as the method presented in this work.

Comments on the overall merit of generating MS^3 data

This paper describes a method for utilizing MS^2 and MS^3 information for cases in which such data has been generated. A fundamental question arises, however, as to whether or not the benefits of generating MS^3 justifies the additional cycle time on the instrument, or whether the additional MS^2 spectra that would be generating in that time would offset the potential advantage. It has recently been suggested (e.g. Ref 43) that the overall benefit of generating MS^3 information for phopsphopeptide experiments may be limited. Although a comprehensive analysis of the merits of MS^3 data generation is beyond the scope of this work, the situation is explored here by comparing sets of mass spectrometry runs on identical samples utilizing both methods: the MS^2/MS^3 cycle discussed above, and an MS^2-only method.

LC-MS/MS analysis was performed on two additional IMAC-enriched whole-cell *D. melanogaster* tryptic digests using a Thermo LTQ, as described in Experimental Procedures. Each sample was separated into two equal fractions which were run individually using the MS^2/MS^3 run method or the MS^2-only method. MS^2 and MS^3 peaklists were extracted from the raw data file and searched separately using SEQUEST. Final SEQUEST reports were then combined into two final result sets for each pair of experiments, one set for the MS^2/MS^3, and one for the MS^2-only data. These four result sets were then analyzed using Peptide/ProteinProphet.

To compare results at both the peptide and protein levels, individual identifications for each of the two final result sets were grouped based either on unique peptide sequence or protein accession numbers. The union, intersection and differences between the MS^2/MS^3 and MS^2-only runs were calculated. The results are displayed as Venn diagrams in Figure 7 for both pairs of experiments. Given that there was significant variation between the number of peptide and protein identifications of the same run method, the two pairs of experiments were not combined to reduce the effect of instrument sampling rate variability in peptide identification, providing a more fair assessment of differences between the two methods. The top pair of Venn diagrams indicate the number of unique proteins identified by each method. Proteins were included in a set if they participated in an identified protein group (see Ref 33) with a group probability of at least 0.95. Proteins from the same group (indistinguishable proteins given the sequences of identified peptides) were counted as a single entry. The lower set of Venn diagrams shows unique peptide identifications. Peptides were included in these sets if their modified sequences were unique, i.e. two peptides with any modification or sequence differences were considered two unique peptides for the main figure. PeptideProphet probability scores of 0.95 or above were required for inclusion. Peptide uniqueness can be defined by a number of standards, however, and the number of identifications listed in each area of the Venn diagram may be overestimated depending on

the definition. The break-out boxes for each of the peptide sets indicate the number for each region of the Venn diagram under four alternative definitions of peptide uniqueness. Under the Type 1 definition, peptides identified from consecutive MS^2 and MS^3 scans that differ only by the loss of one or more phosphate groups on one of the residues (i.e., MS^3 was triggered on the neutral loss) were considered identical and counted as one. Under the Type 2 definition, peptides which differ at the N- or C-terminus by one or more amino acid residues (e.g., due to a missed cleavage) were considered identical, e.g.

```
FVS+80EGDGGHVKPTTF
FVS+80EGDGGHVKPTTFTMR
FVS+80EGDGGHVKPTTFTMRD
```

where S+80 indicates a phosphorylated Ser residue. Under the Type 3 definition, peptides were counted as identical if they had the same sequence but the modification site was ambiguous (residues identified as being phosphorylated are within three amino acid sequences of each other), e.g.

```
KES+80NSEDELEYDPSLYPQR
KESNS+80EDELEYDPSLYPQR
```

Under the Type 4 definition, peptides were counted as unique based on the sequence alone, e.g.:

```
KKES+80NS+80EDELEYDPSLYPQR
KKES+80NSEDELEYDPSLYPQR
KKESNS+80EDELEYDPSLYPQR
KKESNS-18EDELEYDPSLYPQR
```

were considered identical sequences. While these four definitions do not include all possible types and permutations that occur, using them to count peptides allows a more comprehensive comparison between the data sets.

The results indicate that for these data there are potential advantages to both techniques. At the protein level, the majority of proteins were identified by both methods. However, in one pair of runs the MS^2-only method outperformed the MS^2/MS^3 method by identifying 42 more unique proteins than the MS^2/MS^3 method. At the peptide level, the MS^2/MS^3 method was able to identify more phosphorylated peptide forms in both sets of runs

under most of the criteria in which modifications were considered unique (Types 1-3). In terms of the number of unique peptides identified by sequence-alone (Type 4), not taking into account modification state, the MS^2-only set identifies more peptides in one of the runs. This suggests that, at least for certain conditions, sequence coverage may be better with the MS^2-only method.

Overall, these results indicate that generation of MS^3 data may result in a decrease in the number unique peptide and protein identifications. However, several additional comments are necessary for more objective evaluation of the benefits of acquiring MS^3 data. First, the probabilities used in the comparison presented above (Figure 7) were the original probabilities generated by the PeptideProphet and ProteinProphet tools. The probability correction procedure described in this work should permit the selection of a greater number of peptides (and therefore proteins) at a fixed FDR, which would potentially mitigate the loss of sequence coverage. Furthermore if the goal of the study is to identify as many unique modification states as possible, MS^3 data may improve the results. It should also be mentioned that the phosphopeptide data sets used in this work were of high quality (high degree of phosphopeptide enrichment), resulting in sufficiently strong intensity MS signal of phosphopeptide ions and relatively good MS^2 fragmentation. On the other hand, it is possible that in other data sets (e.g., no or poor phosphopeptide enrichment), the relatively low abundance of phosphorylated peptides would lead to less intense MS signal and less interpretable MS^2 spectra, thus making benefits of acquiring MS^3 data more apparent.

Concluding Remarks

The generation of MS^3 information is common in directed areas of proteomics such as phosphopeptide identification. Whether generation of MS^3 information is the best strategy or not is partially dependant on the overall goals of the experiment. Data generated from a

complex phosphopeptide-enriched sample suggest that generation of MS^3 spectra can potentially result in an increased number of unique phosphorylation site identifications. On the other hand, the cycle time spent on generation of MS^3 data does appear to detract from the overall number of unique peptides (by sequence only) and proteins identified in such an experiment. Also, although MS^2 spectra in which neutral loss peaks are dominant are still observed in current generation trap instruments, these spectra appear to frequently contain better backbone fragmentation than older equivalents due to increased ion capacity of the trap. Nevertheless, in experiments in which MS^3 data have been generated, MS^2/MS^3 matching information from the entire experiment can be used to adjust the probabilities of the individual peptide assignments, which has the effect of compensating for the reduced number of MS^2 spectra.

In cases in which a very high certainty in a mapped phosphorylation site is needed, MS^3 experiments are highly valuable as exemplified in the mapping of phosphorylation sites for which biological follow-up experiments are performed. Also, in cases in which neither measurement time nor the amount of phosphopeptide samples are limiting factors, the measurement of MS^3 spectra is advantageous. In fact, in an experimental setup which aims to maximize the number of identified phosphorylation sites from a complex sample, one efficient strategy is to first perform MS^2 experiments and then target specifically the unidentified phosphopeptide ions using MS^2/MS^3 measurements (44, 45).

Generally speaking, much of proteomics data analysis relies on the scores and probabilities produced by automated search algorithms. It is thus important that any probability measure is accurate, and makes use of all available information, particularly in situations where the targeted peptide identifications are rare, e.g. for phosphopeptides and/or when proteins are identified by a reduced number of peptides (such as an analysis in which N-terminal peptides are enriched). Here we have described methods for translating the additional information obtained by matching coupled peptide assignments to MS^2 and MS^3

spectra into a combined probability score, improving the ability to discriminate between true positive and false positive identifications. We have demonstrated an increase in sensitivity and a corresponding decrease in the error rate of selecting correct identifications as a result of the adjusted probability using a mixture of known standard proteins, and applied the method to a complex phosphopeptide-enriched data set, demonstrating an improved discrimination between correct and incorrect peptide assignments for that sample.

The goal of this study was to describe a relatively simple but valid mechanism for adjusting probabilities of peptide identifications in scenarios in which standard database searching has been performed on MS^2/MS^3 data sets. An alternative computational strategy for accommodating MS^3 information is to merge MS^2 and MS^3 spectra into a single spectrum prior to database searching. Full investigation of the relative merits of pre-database search, spectral merging approaches versus a post-database search probability adjustment procedure such as the one discussed here is beyond the scope of this work, but is the subject of current investigation. Other methodologies, such as merging spectra from differently charged precursors of the same peptide, could likely be utilized to improve peptide identification as well.

As instrumentation continues to improve the speed and accuracy of tandem MS measurements, the ability to generate complementary information such as MS^3 spectra for any given ion will become increasingly practical. Methods for accommodating this information are consequently useful, and can significantly improve the quality of the results generated by automated processing of mass spectrometry data.

Data and Code Availability

MzXML and RAW datafiles, and processed unique linked pair data, for both the 9-Mix and phospho samples are available online via the Tranche system (http://tranche.proteomecommons.org). The software used in this work was developed in Python. Python modules were implemented making use of the code library available with the InsPecT software package by the UCSD Computational Mass Spectrometry Research Group (28). All code modules generated by our group for this project are available upon request.

Acknowledgements

This work was supported in part by NIH/NCI Grant CA-126239 to AIN, NIH/NCRR - National Resource for Proteomics and Pathways Grant #P41-18627 to PCA, and with funds from NIH/NHLBI under contract No. N01-HV-28179 to RA. Bernd Bodenmiller is the recipient of a fellowship by the Boehringer Ingelheim Fonds. We thank Steven Tanner and the UCSD Computational Research Group for the free availability of their code. All annotated spectra in this manuscript were generating using the Label.py and MakeImage.py modules available in the InsPecT library.

References

1. Hager, J. W. (2002) A new linear ion trap mass spectrometer. *Rapid Commun in Mass Spectrom* 16, 512-526
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198-207
3. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem Rev* 101, 269-295
4. Nesvizhskii, A. I. (2006) Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* 367, 87-120
5. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 4, 1419-1440
6. Sadygov, R. G., Cociorva, D., and Yates, J. R., 3rd (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 1, 195-202
7. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5, 699-711
8. Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes J. S., Gruissem, W., Baginsky, S., Aebersold, R. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* 5, 652-70.
9. Pevzner, P. A., Mulyukov, Z., Dancik, V. and Tang, C. L. (2001) Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Res.* 11: 290-299.
10. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V., and Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol* 7, R80
11. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 101, 13417-13422
12. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 101, 12130-12135
13. Bodenmiller, B., Mueller, L. N., Pedrioli, P. G. A., Pflieger, D., Jünger, M. A., Eng, J., Aebersold, R., and Tao, W. A. (2007) An integrated chemical, mass spectrometric and computational strategy for (quantitative) phosphoproteomics: Application to Drosophila melanogaster Kc167 Cells. *Molecular BioSystems*, 3, 275-286.
14. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 4, 310-327
15. Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol Cell Proteomics* 5, 949-958
16. Zabrouskov, V., Senko, M. W., Du, Y., Leduc, R. D., and Kelleher, N. L. (2005) New and automated MSn approaches for top-down identification of modified proteins. *J Am Soc Mass Spectrom* 16, 2027-2038
17. Zhang, Z., and McElvain, J. S. (2000) De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal Chem* 72, 2337-2350
18. Demelbauer, U. M., Zehl, M., Plematl, A., Allmaier, G., and Rizzi, A. (2004) Determination of glycopeptide structures by multistage mass spectrometry with low-energy

collision-induced dissociation: comparison of electrospray ionization quadrupole ion trap and matrix-assisted laser desorption/ionization quadrupole ion trap reflectron time-of-flight approaches. *Rapid Commun Mass Spectrom* 18, 1575-1582
19. LeDuc, R. D., Taylor, G. K., Kim, Y. B., Januszyk, T. E. Bynum, L. H., Sola, J. V., Garavelli, J. S., Kelleher, N. L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* 32, W340-W345
20. Frank, A., and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 77, 964-973
21. Goodlett, D. R., Keller, A., Watts, J. D., Newitt, R., Yi, E. C., Purvine, S., Eng, J. K., von Haller, P., Aebersold, R., and Kolker, E. (2001) Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun Mass Spectrom* 15, 1214-1221
22. Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res* 11, 290-299
23. Regnier, F. E., Liu, P. (2002) An isotope coding strategy for proteomics involving both amine adn carboxly group labeling. *J Proteome Res*,1, 443-50
24. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551-3567
25. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467
26. Eng, J. K., McCormack, A. L., and Yates, J. R. r. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976-989
27. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3, 958-964
28. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 77, 4626-4639
29. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2, 1406-1412
30. Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., and Aebersold, R. (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nature methods* 4, 231-237
31. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1, 2005.0017
32. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74, 5383-5392
33. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75, 4646-4658
34. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33:D154-159.

35. Salek, M., Lehmann, W. D. (2003) Neutral loss of amino acid residues from protonated peptides in collision-induced dissociation generates N- or C-terminal sequence ladders. *J Mass Spectrom* 38, 1143-9
36. Martin, D. B., Eng, J. K,, Nesvizhskii, A. I., Gemmill, A., Aebersold, R. (2005) Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal Chem* 77, 4870-82
37. Malmstrom, J., Lee, H., Nesvizhskii, A. I., Shteynberg, D., Mohanty, S., Brunner, E., Ye, M., Weber, G., Eckerskorn, C., and Aebersold, R. (2006) Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J Proteome Res* 5, 2241-2249
38. MacCoss, M. J., Wu, C. C., Yates, J. R. 3rd. (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem.* 74, 5593-9
39. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635-648
40. Villen, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* 104, 1488-1493
41. Elias, J. E., Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. Mar 4: 207-14..
42. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2, 43-50
43. Li, X., Gerber, S. A., Rudner, A. D., Beausoleil, S. A., Haas, W., Villen, J., Elias, J. E., Gygi, S. P. (2007) Large-Scale Phosphorylation Analysis of alpha-Factor-Arrested Saccharomyces cerevisiae. *J Proteome Res* 6, 1190-7
44. Picotti, P., Aebersold, R., Domon, B. (2007) The Implications of Proteolytic Background for Shotgun Proteomics. *Mol Cell Proteomics*. May 28 [Epub ahead of print]
45. Domon, B. Aebersold, R. (2006) Mass spectrometry and protein analysis. Science 312, 212-7.

Figure Legends

Figure 1. Overview of methodology. MS^2 and MS^3 spectra are extracted from the raw data and the spectra are assigned peptides using sequence database searching (SEQUEST or similar programs). The resulting peptide assignments are statistically validated using PeptideProphet, which calculates for each assignment in the data set a probability of being correct (applied separately for MS^2 and MS^3 data). MS^2 and MS^3 scan results are correlated based on scan number, in which an MS^3 spectrum is linked to an MS^2 if its scan number is consecutive. Based on the overall matched data set, a Bayesian probability correction is applied to linked scan results individually for MS^2 and MS^3 spectra, resulting in adjusted probability scores. In the final step, the MS^2 and MS^3 scan results are combined and a final probability calculated for each scan number as representative of the peptide identification.

Figure 2. Examples of MS^2/MS^3 linked pairs and the probability correction procedure. MS^2 (left) and matching consecutive MS^3 peaklists (right) are shown. The charge state of each spectrum is indicated in the upper left corner. a) Example of the probability correction for a +3 MS^2 -> +2 MS^3 matched pair. b) A +2/+1 match pair for a phosphopeptide identification in which the y_{12} ion is selected for MS^3. c) A +3/+1 identification; the y_8 ion is selected for MS^3. d) An example of a +2/+2 loss of the phosphate moiety in which the most abundant MS^2 peak selected for MS^3 is the doubly-charged y_{13} – 98 Da.

Figure 3. Performance of MS^2 and MS^3 scores with probability adjustment. Error rate of MS^2 a) and MS^3 b) scores are shown as a function of sensitivity for initial (dashed) and adjusted (solid) probabilities. Inserted panels are zoomed areas of the plots for the 0 - 10% error rate range.

Figure 4. Discriminating power and accuracy of computed probabilities. a) Total number of correct peptide assignments is plotted as a function of minimum probability threshold for MS^2 and MS^3 spectra alone, both initial and adjusted, and both P_{max} and P_{comb} scores. b) Same as a), zoomed in the region of minimum probability threshold 0.9 to 1.0. c) Number of correct peptide assignments as a function of the number of incorrect assignments, plotted separately for MS^2 (green) and MS^3 (blue) initial (dashed) and adjusted (solid) probabilities,

as well as the combined P_{max} (red) and P_{comb} (purple). d) Probability accuracy of the adjusted MS^2, MS^3, P_{max} and P_{comb} probabilities.

Figure 5. Performance of probability scores on the phosphopeptide data set. The number of correct identifications estimated using the decoy database method is plotted as a function of FDR estimated using the decoy database search method. a) Results for the phosphopeptide data set, and b) for the non-phosphorylated identifications only in the phosphopeptide data set. For MS^2 and MS^3 results, dashed lines indicate initial and solid lines indicate corrected probability scores.

Figure 6. Degree of probability score adjustment by sequence match category for the 9-Mix and phosphopeptide data sets.

Figure 7. Comparison of MS^2/MS^3 and MS^2-only experimental runs. Two equivalent pairs of runs are shown, labeled Run1 – Run4. Venn diagrams display overlap between MS^2/MS^3 (left) and MS^2-only (right) data sets based on unique identifications at the peptide and protein levels. All identifications are based on a 95% probability threshold. The top diagrams display protein identifications based on unique Uniprot entry name. The numbers represent the number of ProteinProphet protein groups that have a protein group probabilities equal or greater than 0.95. The lower diagrams show the same for peptide identifications based on peptide sequence, using initial PeptideProphet probability scores. Peptide identifications with the least stringent, most inclusive uniqueness criteria are shown in the main figure. Counts for each region of the diagram utilizing more stringent uniqueness criteria are shown in the boxes, labeled as "- Type".

Table 1. Results of consecutive MS^2/MS^3 scan pairing for the 9-Mix data set. a) Overall pair counts. Total MS^2/MS^3 peptide assignments shows the number of identifications of the indicated type in the data set. Number of unique pairs indicates the number of pairs generated by linking consecutive scan numbers and selecting the most likely pair amongst all possibilities. MS^2 search results without consecutive MS^3 results are still counted in these numbers. Panel b) shows the number of unique pairs for the MS^2 and unique matching sets after binning into sequence match categories, as described in the text. Panel c) shows truth category bin counts of the MS^2 and unique matching sets. A '+' in the truth category column descriptors indicates a correct match, '-' an incorrect, and 'null' the lack of a consecutive MS^3 for an MS^2 scan.

Table 1. Consecutive MS^2/MS^3 scan pairing for the 9-Mix dataset

a | Overall pair counts

Total MS^2 peptide assignments:	**48921**
Total MS^3 peptide assignments:	**27952**
Number of unique pairs:	**23367**
Number of (non-null) MS^3 peptide assignments in unique pair set:	**16140**

b | Sequence match category counts

	0: No Consecutive	**1**:Consecutive (no match)	**2**: MS^3 Seq In MS^2 Seq	**3**:MS^3 Seq = MS^2 Seq	**4**: MS^2 Seq In MS^3 Seq
Unique Matches	**7227**	**14665**	**1458**	**0**	**17**

c | Truth category counts

	+/null	-/null	+/+	+/-	-/+	-/-
Unique Matches	**1025**	**6202**	**1509**	**2029**	**1287**	**11315**

Table 2. Posterior probabilities of observing a correctly (+) or incorrectly (-) assigned peptide to a MS^2 or MS^3 scan among peptides from the four most frequently observed sequence match categories in the 9-Mix data set (0: no consecutive scan; 1: consecutive scan, no matching sequence; 2: consecutive scan, MS^3 sequence is a subset of MS^2 sequence; 3: consecutive scan, MS^3 sequence identical to MS^2 sequence).

Table 2 | Posterior Match Probabilities for the 9-Mix dataset

	Sequence Match Category			
	0	**1**	**2**	**3**
MS^2 *p(Match|+)*	0.213	0.483	0.304	0.000
MS^2 *p(Match|-)*	0.332	0.662	0.005	0.000
MS^3 *p(Match|+)*	0.000	0.584	0.416	0.000
MS^3 *p(Match|-)*	0.000	0.960	0.040	0.000

Table 3. Statistics for the phosphopeptide-enriched data set. a) Total match counts; b) Sequence match category counts; c) Posterior probabilities for the statistical model. Sequence match category numbers are described in the caption for Table 2.

Table 3. Consecutive MS^2/MS^3 scan pairing for the Phosphopeptide-enriched dataset

a | Overall pair counts

Total MS^2 peptide assignments:	**16647**
Total MS^3 peptide assignments:	**12218**
Number of unique pairs:	**7547**
Number of (non-null) MS^3 peptide assignments in unique pair set:	**6270**

b | Sequence match category counts

	0: No Consecutive	**1**:Consecutive (no match)	**2**: MS^3 Seq In MS^2 Seq	**3**:MS^3 Seq = MS^2 Seq
Unique Matches	**1277**	**5788**	**167**	**313**

c | Posterior Match Probabilities

	Sequence Match Category			
	0	**1**	**2**	**3**
MS^2 *p(Match|+)*	0.086	0.676	0.079	0.159
MS^2 *p(Match|-)*	0.196	0.796	0.004	0.004
MS^3 *p(Match|+)*	0.000	0.424	0.131	0.445
MS^3 *p(Match|-)*	0.000	0.980	0.015	0.005

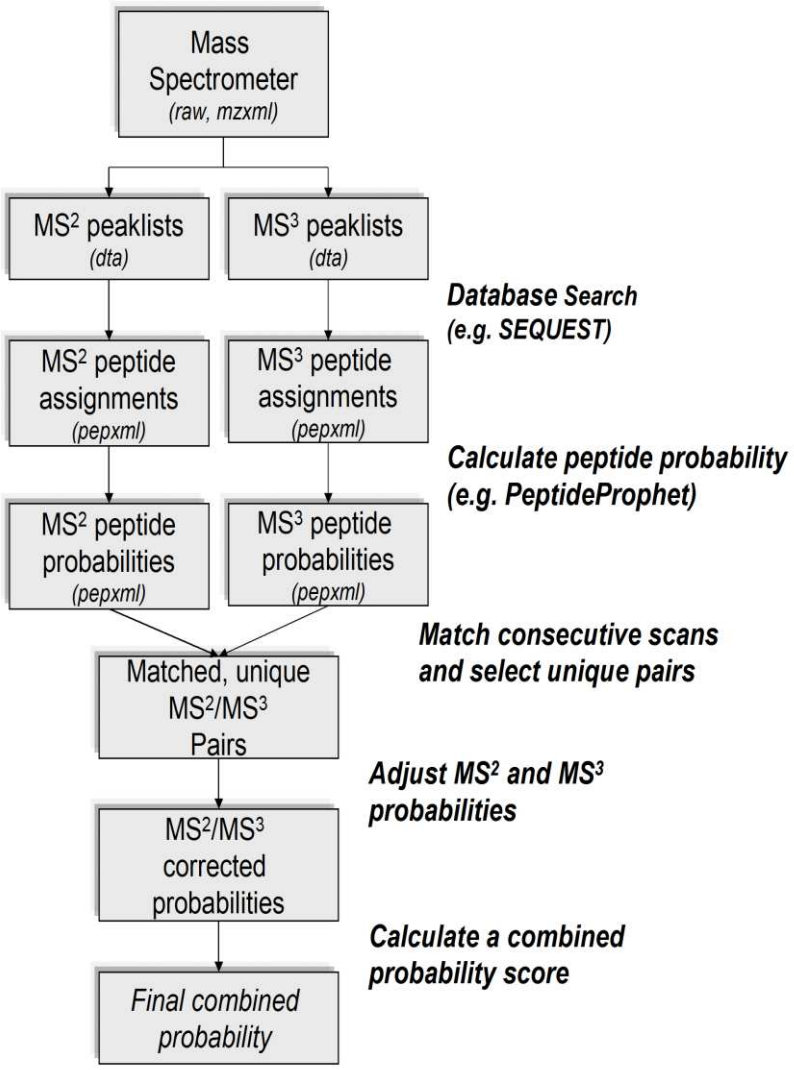
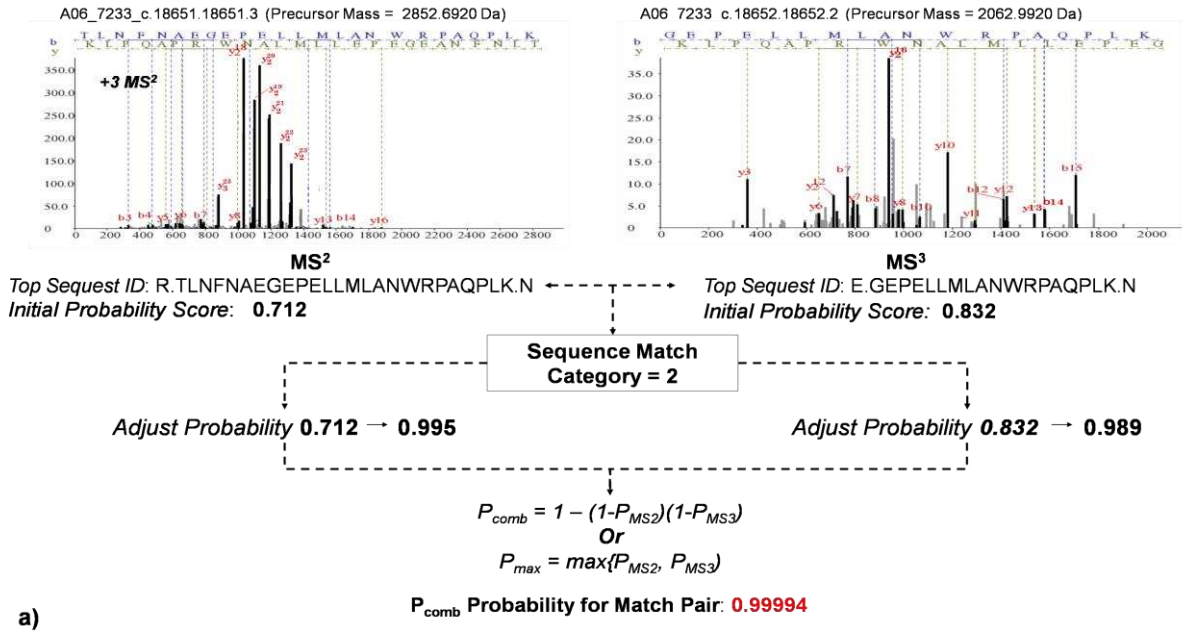
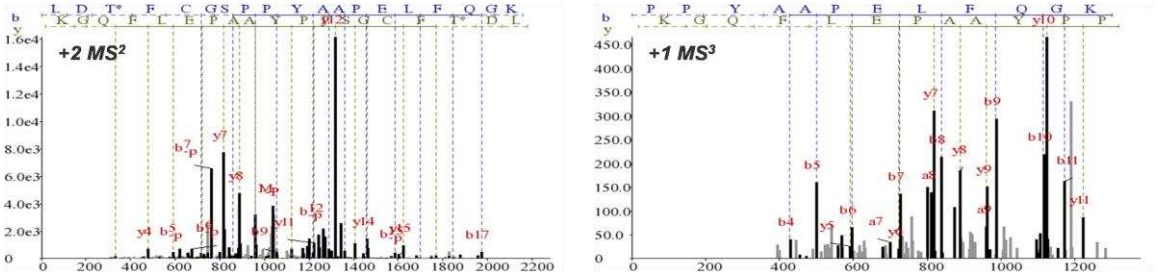

Mass Spectrometer
(raw, mzxml)
MS2 peaklists
(dta)
MS3 peaklists
(dta)
Database Search
(e.g. SEQUEST)
MS2 peptide assignments
(pepxml)
MS3 peptide assignments
(pepxml)
Calculate peptide probability
(e.g. PeptideProphet)
MS2 peptide probabilities
(pepxml)
MS3 peptide probabilities
(pepxml)
Match consecutive scans
and select unique pairs
Matched, unique MS2/MS3 Pairs
Adjust MS2 and MS3 probabilities
MS2/MS3 corrected probabilities
Calculate a combined probability score
Final combined probability

Figure 1

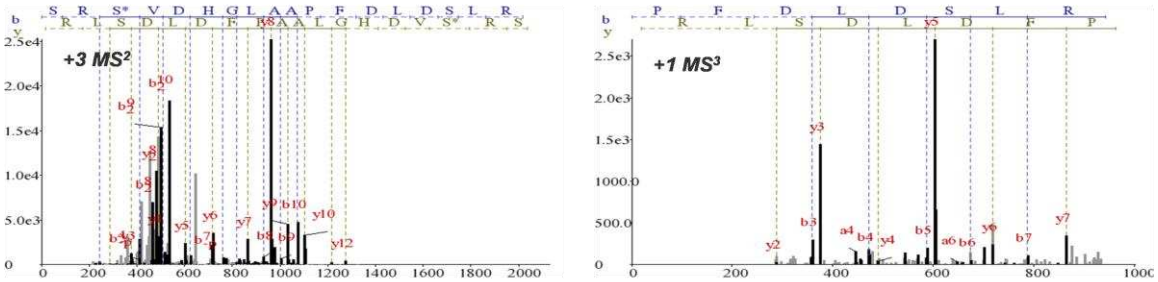


a)



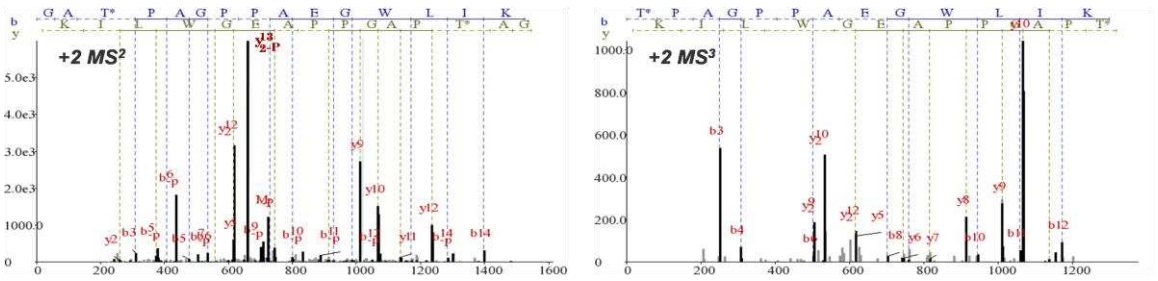
b) A06_6009_p.18637.18637.2 Precursor = 2177.58 Da

A06_6009_p.18638.18638.1 Precursor = 1317.58 Da



c) A06_6009_p.16742.16742.3 Precursor = 2036.06 Da

A06_6009_p.16743.1743.1 Precursor = 961.82 Da



d) A06_6009_p.14679.14679.2 Precursor = 1544.64 Da

A06_6009_p.14680.14680.2 Precursor = 660.12 Da

Figure 2

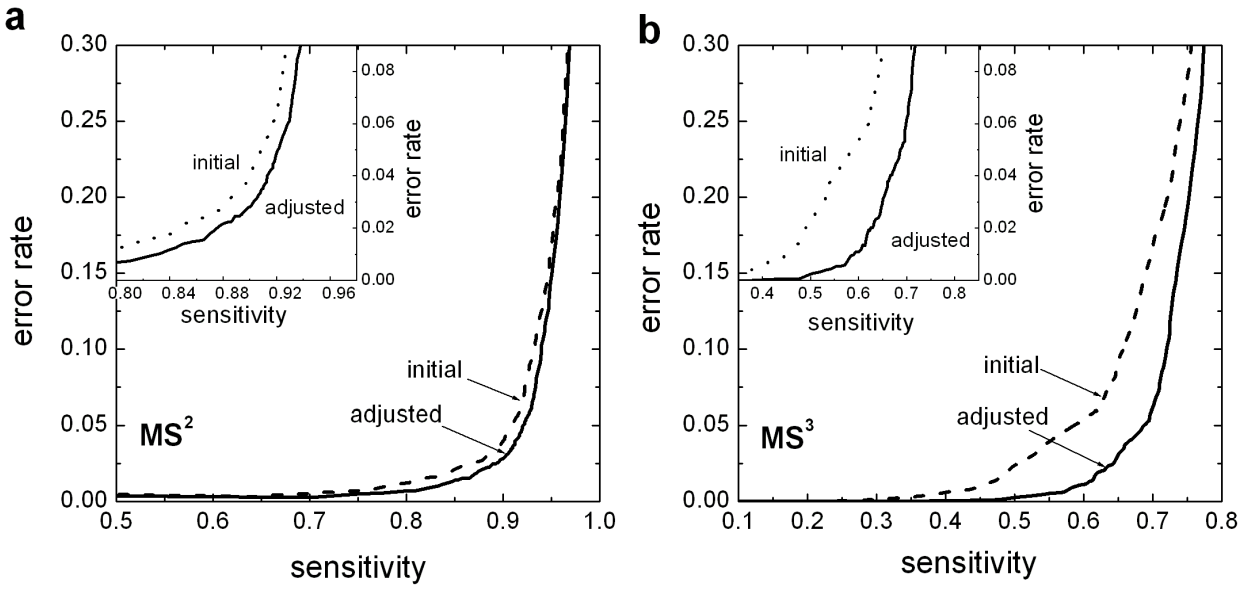

a
b
error rate
sensitivity
initial
adjusted
MS2
MS3
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.5
0.6
0.7
0.8
0.9
1.0
0.1
0.2
0.3
0.4
0.08
0.06
0.04
0.02
0.00
0.80
0.84
0.88
0.92
0.96

Figure 3

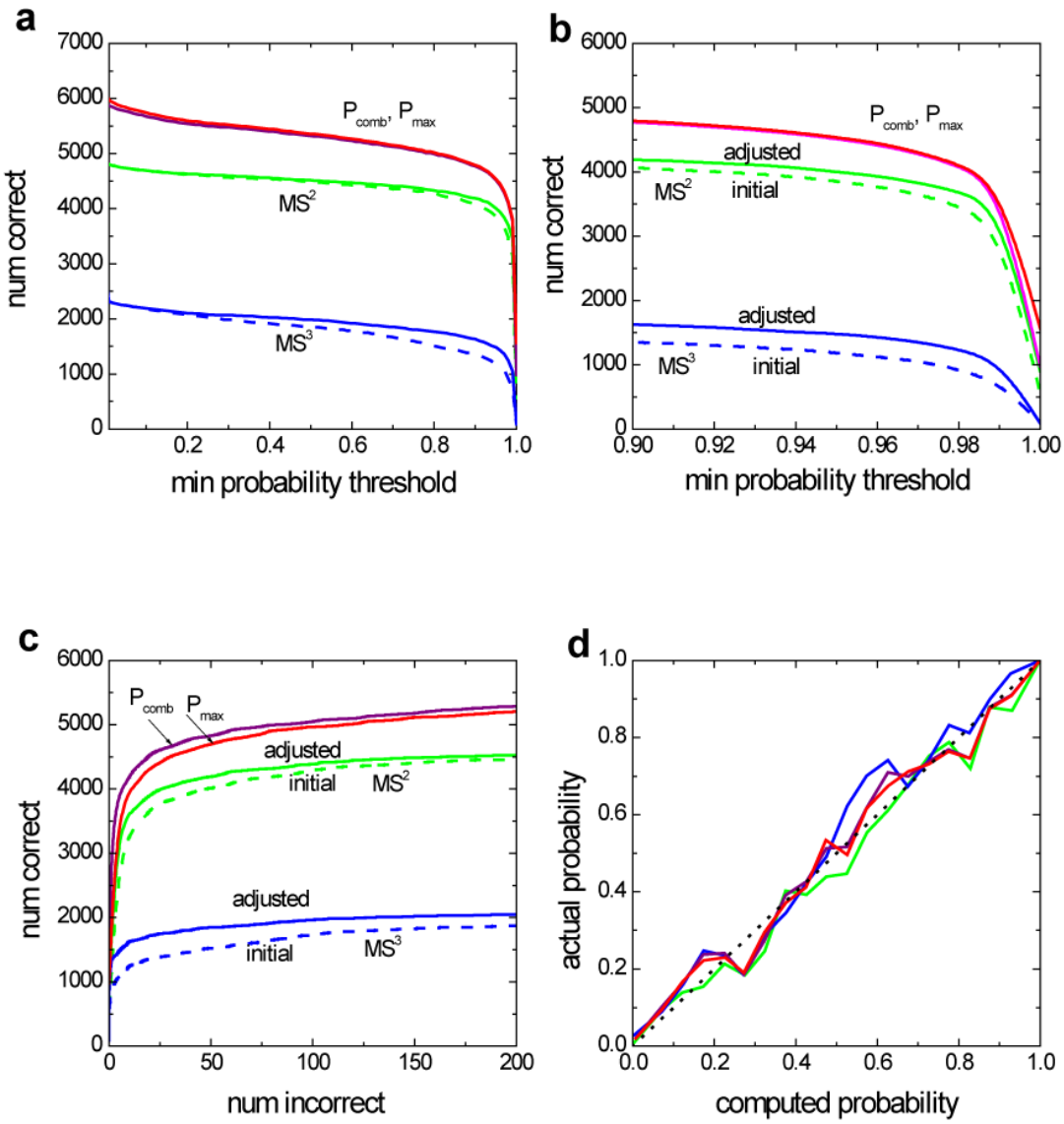

a
num correct
min probability threshold
P_comb, P_max
MS2
MS3
b
num correct
min probability threshold
P_comb, P_max
adjusted
initial
MS2
MS3
c
num correct
num incorrect
P_comb
P_max
adjusted
initial
MS2
MS3
d
actual probability
computed probability

Figure 4

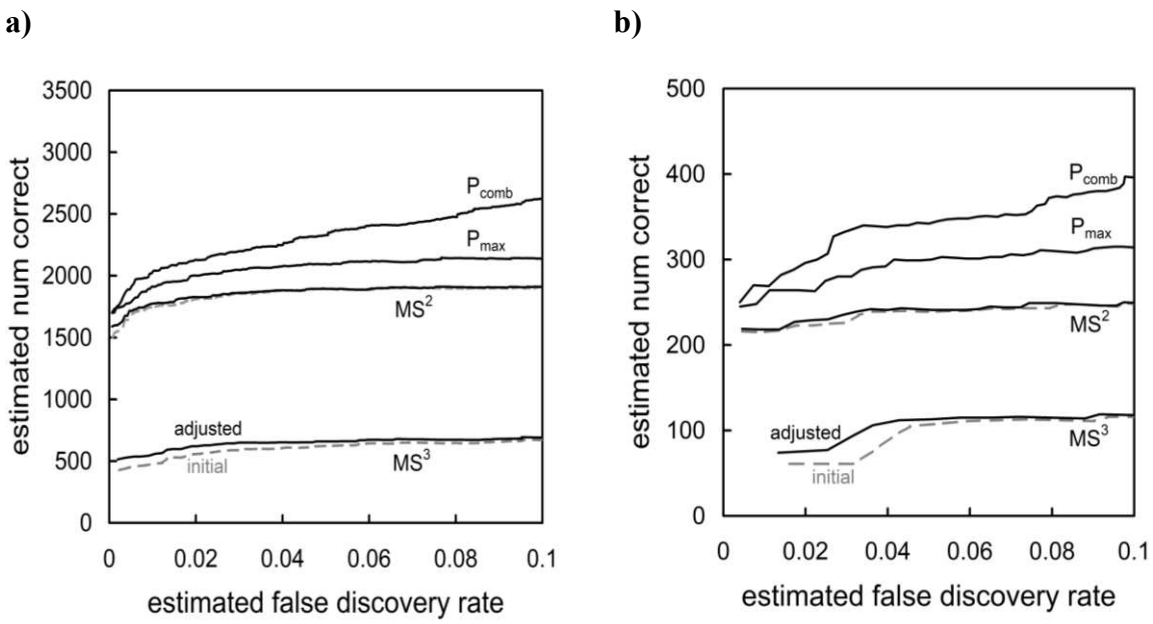

a)
b)
estimated num correct
estimated false discovery rate
P comb
P max
MS2
MS3
adjusted
initial
0
0.02
0.04
0.06
0.08
0.1
500
1000
1500
2000
2500
3000
3500
100
200
300
400

Figure 5

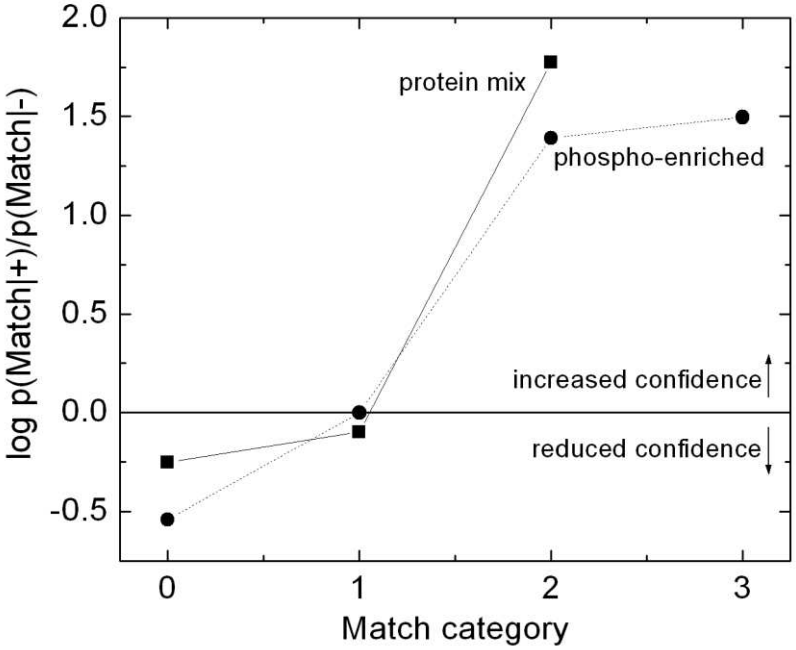

2.0
1.5
1.0
0.5
0.0
-0.5
log p(Match|+)/p(Match|-)
protein mix
phospho-enriched
increased confidence
reduced confidence
0
1
2
3
Match category

Figure 6

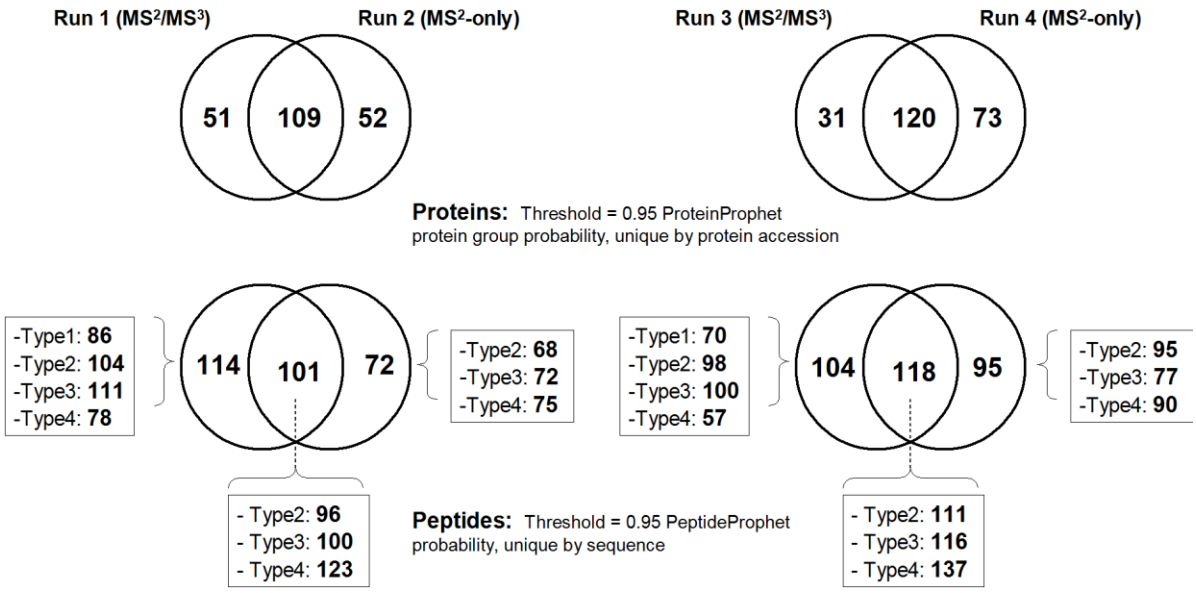

Run 1 (MS²/MS³)
Run 2 (MS²-only)
51
109
52
Run 3 (MS²/MS³)
Run 4 (MS²-only)
31
120
73
Proteins: Threshold = 0.95 ProteinProphet protein group probability, unique by protein accession
-Type1: 86
-Type2: 104
-Type3: 111
-Type4: 78
114
101
72
-Type2: 68
-Type3: 72
-Type4: 75
- Type2: 96
- Type3: 100
- Type4: 123
-Type1: 70
-Type2: 98
-Type3: 100
-Type4: 57
104
118
95
-Type2: 95
-Type3: 77
-Type4: 90
- Type2: 111
- Type3: 116
- Type4: 137
Peptides: Threshold = 0.95 PeptideProphet probability, unique by sequence

Figure 7